CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY
FY 2016 THIRD QUARTER REPORT
–April 2016 through June 2016 –

COLLABORATION

DECENNIAL DIRECTORATE

Decennial Management Division/Decennial Statistical Studies Division/American Community Survey Office (Sponsors)

Project Number	Project Title	FTEs
6650B23	Redesigning Field Operations	1.32
6750B01	Administrative Records Data	3.87
6550B01	Data Coding, Editing, and Imputation	0.15
6250B07	Policy	0.25
	A. Decennial Record Linkage	
	B. Coverage Measurement Research	
	C. Using 2010 Census Coverage Measurement Data to Compare Nonresponse Follow-up Proxy Responses with Administrative Records	
	D. Record Linkage Error-Rate Estimation Methods	
	E. Supplementing and Supporting Non-Response with Administrative Records	
	F. Identifying "Good" Administrative Records for 2020 Census NRFU Curtailment Targeting	
	G. Evaluation of Response Error Using Administrative Records	
	H. Special Census: Disclosure Avoidance in Group Quarters	
	I. 2020 Unduplication Research	
6350B02	Address Canvassing in Field	1.59
	A. Master Address File (MAF) Error Model and Quality Assessment	
	B. Development of Block Tracking Database	
	C. Detection of Map Changes	
6385B70	American Community Survey (ACS)	2.72
	A. ACS Applications for Time Series Methods	
	B. ACS Imputation Research and Development	
	C. Data Analysis of ACS CATI-CAPI Contact History	
	D. Assessing Uncertainty in ACS Ranking Tables	
	E. Confidence Intervals for Proportions in ACS Data	
	F. Voting Rights Section 203 Model Evaluation and Enhancements Towards Mid-Decadal Determinations	

DEMOGRAPHIC DIRECTORATE

Demographic Statistical Methods Division (Sponsor)

Project Number	Project Title	FTEs
TBA	Demographic Statistical Division Special Projects	TBA
	A. Special Project on Weighting and Estimation	

Demographic Surveys Division (Sponsor)

Project Number	Project Title	FTEs
0906/1444X00	Demographic Surveys Division Special Projects	0.70
	A. Data Integration	

Population Division (Sponsor)

Project Number	Project Title	FTEs
TBA	Population Division Projects	TBA
	A. Introductory Sampling Workshop	

Social, Economic, and Housing Statistics Division (Sponsor)

Project Number	Project Title	FTEs
7165016	Social, Economic, and Housing Statistics Division Small Area Estimation Projects	2.38
	A. Research for Small Area Income and Poverty Estimates (SAIPE)	
	B. Small Area Health Insurance Estimates (SAHIE)	
	C. Sub County Estimates of Poverty from Multi-year ACS Data	

ECONOMIC DIRECTORATE

Project Number	Project Title	FTEs
1183X01	Economic Statistical Collection	0.25
1001X00	Economic Monthly/Retail	0.06
	A. Research on Imputation Methodology for the Monthly Wholesale Trade Survey	
	B. Use of Big Data for Retail Sales	
2220B10	Economic Census/Survey Engineering: Time Series Research; Economic Missing Data/Product Line Data; Development/SAS	3.25
	A. Seasonal Adjustment Support	
	B. Seasonal Adjustment Software Development and Evaluation	
	C. Research on Seasonal Time Series - Modeling and Adjustment Issues	
	D. Supporting Documentation and Software for X-13ARIMA-SEATS	
	E. Missing Data Adjustment Methods for Product Data in the Economic Census	
7103012	2012 Commodity Flow Survey	0.03
	A. Commodity Flow Survey	
TBA	Investigation of Alternative Methods for Resolving Balance Complex Failures in StEPS	TBA
	A. Investigation of Alternative Methods for Resolving Balance Complex Failures in StEPS	

RESEARCH AND METHODOLOGY DIRECTORATE

Center for Economic Studies (Sponsor)

Project Number	Project Title	FTEs
TBA	Business Dynamics Statistics—Export Weighting Issue	TBA
	A. Business Dynamics Statistics—Export Weighting Issue	

CENSUS BUREAU

Project Number	Project Title	FTEs
7236045	National Survey of Drug Use & Health	0.06
	A. National Survey of Drug Use & Health	
0331000	Program Division Overhead	9.00
	A. Center Leadership and Support	
	B. Research Computing	

GENERAL RESEARCH AND SUPPORT

Project Number	Project Title	FTEs
0331000	General Research and Support	26.05
0925000	General Research	8.28

MISSING DATA, EDIT, AND IMPUTATION

- *A. Editing*
- *B. Editing and Imputation*

RECORD LINKAGE

- *A. Disclosure Avoidance for Microdata*
- *B. Record Linkage and Analytic Uses of Administrative Lists*
- *C. Modeling, Analysis and Quality of Data*
- *D. R Users Group*

SMALL AREA ESTIMATION

- *A. Small Area Methods with Misspecification*
- *B. Coverage Properties of Confidence Intervals for Proportions in Complex Surveys*
- *C. Small Area Estimates of Disability*
- *D. Using ACS Estimates to Improve Estimates from Smaller Surveys via Bivariate Small Estimation Models*
- *E. Multivariate Fay-Herriot Hierarchical Bayesian Estimation of Small Area Means under Functional Measurement Error*
- *F. Smoothing Design Effects for Small Sample Areas*

SURVEY SAMPLING-ESTIMATION AND MODELING

- *A. Household Survey Design and Estimation*
- *B. Sampling and Estimation Methodology: Economic Surveys*
- *C. The Ranking Project: Methodology Development and Evaluation*
- *D. Sampling and Apportionment*
- *E. Interviewer-Respondent Interactions: Gaining Cooperation*
- *F. Weighted Estimating Equations with Response Propensities in Terms of Covariates Observed Only for Responders*
- *G. Analysis and Estimation of Daily Response Propensities and Use of Contact History Instrument*

TIME SERIES AND SEASONAL ADJUSTMENT

- *A. Seasonal Adjustment*
- *B. Time Series Analysis*
- *C. Time Series Model Development*

EXPERIMENTATION AND STATISTICAL MODELING

- *A. Design and Analysis of Embedded Experiments*
- *B. Synthetic Survey and Processing Experiments*
- *C. Multivariate Nonparametric Tolerance Regions*
- *D. Master Address File (MAF) Research—Developing a Generalized Regression Model for Count Data*
- *E. Modeling the Causal Effects of Field Representative Actions and Strategies*
- *F. Development of a Bivariate Distribution for Count Data where Data Dispersion is Present*
- *G. Developing a Flexible Stochastic Process for Significantly Dispersed Count Data*
- *H. Analysis of Under-dispersed Count Data*

SIMULATION AND STATISTICAL MODELING

- *A. Development and Evaluation of Methodology for Statistical Disclosure Control*

SUMMER AT CENSUS

RESEARCH SUPPORT AND ASSISTANCE

PUBLICATIONS

- Journal Articles, Publications
- Books/Book Chapters
- Proceedings Papers
- Center for Statistical Research & Methodology Research Reports
- Center for Statistical Research & Methodology Studies

TALKS AND PRESENTATIONS

CENTER FOR STATISTICAL RESEARCH & METHODOLOGY SEMINAR SERIES

PERSONNEL ITEMS

- Honors/Awards/Special Recognition
- Significant Service to Profession
- Personnel Notes

1. COLLABORATION

1.1 REDESIGNING FIELD OPERATIONS (Decennial Project 6650B23)

1.2 ADMINISTRATIVE RECORDS DATA (Decennial Project 6750B01)

1.3 DATA CODING, EDITING, AND IMPUTATION (Decennial Project 6550B01)

1.4 POLICY (Decennial Project 6250B07)

A. Decennial Record Linkage
Description: Under this project, staff will provide advice, develop computer matching systems, and develop and perform analytic methods for adjusting statistical analyses for computer matching error with a decennial focus.

Highlights: Staff wrote a summary related to the accuracy and speed of *BigMatch* that was forwarded to 20+ individuals. In a three-year review by professors at Curtin University in Western Australia, *BigMatch* was considered the most accurate software in comparison with commercial products from IBM and SAS and four shareware products written by university professors for the health agencies. *BigMatch* is 50 times as fast as parallel software from CS professors at Stanford and 10 times as fast as recent parallel software written by three other sets of CS professors. Staff provided extensive advice and software to individuals at Carnegie-Mellon University who are working on a grant funded by the Decennial Statistical Studies Division (DSSD). Staff provided advice and details on some of the computational algorithms in *BigMatch*.

Staff: William Winkler (x34729), Emanuel Ben-David

B. Coverage Measurement Research
Description: Staff members conduct research on model-based small area estimation of census coverage, and they consult and collaborate on modeling census coverage measurement (CCM).

Highlights: During Q3 of FY 2016, staff attended weekly meetings to discuss and help make decisions on how the 2020 Census Coverage Measurement program will be conducted.

Staff: Jerry Maples (x32873), Ryan Janicki, Eric Slud

C. Using 2010 Census Coverage Measurement Data to Compare Nonresponse Follow-up Proxy Responses with Administrative Records

Description: Research in preparation for the 2020 Census Nonresponse Follow-up (NRFU) investigates employing different contact strategies combined with the use of administrative records (AR) files in different ways in order to reduce the cost of the operation while maintaining data quality. Regardless of the contact strategy, the question arises as to whether the proxy responses are more accurate than ARs available for the

NRFU housing units (HUs). The goal of this study is to use the results of the 2010 Census Coverage Measurement Program (CCM) to compare the accuracy of proxy responses for 2010 Census NRFU housing units in the CCM sample with the accuracy of the ARs available for the housing units.

Highlights: Staff received comments on a draft research report and will revise as necessary.

Staff: Mary Mulry (x31759)

D. Record Linkage Error-Rate Estimation Methods
Description: This project develops methods for estimating false-match and false-nonmatch rates without training data and with exceptionally small amounts of judiciously chosen training data. It also develops methods/software for adjusting statistical analyses of merged files when there is linkage error.

Highlights: No significant updates this quarter.

Staff: William E. Winkler (x34729), Emanuel Ben-David, Tom Mule (DSSD)

E. Supplementing and Supporting Non-Response with Administrative Records
Description: This project researches how to use administrative records in the planning, preparation, and implementation of nonresponse follow-up to significantly reduce decennial census cost while maintaining quality. The project is coordinated by one of the 2020 Census Integrated Project Teams.

Highlights: Staff continued to analyze the results of stepwise logistic regression models with top-coded Census Unedited File (CUF) household size as the dependent variable. The models used a national data file of 2010 Decennial Census nonresponse follow-up (NRFU) IDs. Staff fit models on a random 5% subsample and scored the models on the entire file. Staff compared several different models on the group of IDs where the undeliverable as addressed (UAA) flag is

blank and the 2009 1040 household count is at least one. For these IDs, the distribution of the expected value of household size (by top-coded IRS household size) and the distribution of the rounded expected value of household size are reasonably similar across models, although adding variables to a model does tend to increase the spread of the expected value. Separate models for each value of top-coded IRS household size, however, are still needed for the distribution of the predicted value (household size with the maximum estimated probability) to be similar to the distribution of top-coded CUF household size. None of the models has a distribution of rounded expected value of household size that is similar to the distribution of top-coded CUF household size.

Staff also looked at the effect of using cutoffs for the maximum estimated probability in determining when to use the household count based on administrative records. This part of the analysis included IDs where the UAA flag was blank and either the 2009 IRS 1040 household count or the 2008 IRS 1040 household count was at least one. For IDs where the 2009 IRS 1040 household count was at least one, there were separate stepwise logistic regression models for each value of top-coded 2009 IRS 1040 household size. For IDs where the 2009 IRS 1040 household count was zero but the 2008 IRS 1040 household count was at least one, there were separate stepwise logistic models for each value of top-coded 2008 IRS 1040 household size. After both sets of models were fit and scored, the output from the two sets of models was combined, IDs with predicted values other than 1-6 were excluded, and quartiles of the maximum estimated probability were calculated based on the remaining IDs and used as cutoffs. As the cutoff got more restrictive, the rounded expected value and the predicted value became more concentrated on household sizes of one, two, and four (especially size one) and also showed fewer differences from each other. For the predicted value, using no probability cutoff resulted in an overestimate compared to the CUF, while using the third quartile as the cutoff resulted in an underestimate. For the rounded expected value, the estimates are similar to the CUF results both for no cutoff and for cutoffs at any quartile. The results are similar when IDs enumerated on the first contact are ineligible for using the modeled count, although the distributions of the remaining rounded expected values, predicted values, and CUF household counts shift a bit towards smaller counts. The CUF household counts tend to be shifted a bit more on average than the rounded expected values or predicted values. This means that the ratio of the AR-based estimates to the CUF results increases somewhat (compared to the corresponding ratios including IDs enumerated on the first contact) for those units that meet the cutoffs. Because there are fewer of these units, however, the net effects on the NRFU estimates are similar.

Staff: Michael Ikeda (x31756), Mary Mulry

F. Identifying "Good" Administrative Records for 2020 Census NRFU Curtailment Targeting

Description: As part of the Census 2020 Administrative Records Modeling Team, staff are researching scenarios of nonresponse follow-up (NRFU) contact strategies and utilization of administrative records data. Staff want to identify scenarios that have reduction in NRFU workloads while still maintaining good census coverage. Staff are researching identification of "good" administrative records via models of the match between Census and administrative records person/address assignments for use in deciding which NRFU households to continue to contact and which to primary allocate. Staff are exploring various models, methods, and classification rules to determine a targeting strategy that obtains good Census coverage—and good characteristic enumeration—with the use of administrative records.

Highlights: During Q3 of FY 2016, staff finalized the paper "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Decennial Census" published in the June issue of the *Statistical Journal of the International Association of Official Statistics* as part of a series of papers on using administrative records in the 2020 Census. This paper documents a scenario of administrative records vacancy and occupancy (enumeration) determination using a linear programming approach. Staff also completed a peer-reviewed journal revise/resubmit request of a paper comparing classification methods for a person-place model for administrative records usage. This revision incorporated additional background on the use of administrative records in Decennial operations as well as sub-national results of applying the methodology in a retrospective study of the 2010 Census. Staff continued to attend meetings and provide input into research topics studied by the administrative records modeling team such as comparison of models using 2010 Census vs. ACS data, an alternative approach for determining administrative records removals using a distance function based on predicted probabilities, and the analysis of USPS information in the 2016 Census test.

Staff investigated decision theoretical strategies for conducting NRFU by exploiting information on response propensity along with information on administrative records (AR) quality. The decision-theoretical strategies optimize field effort in concert with administrative record substitution to identify the most productive combination of fieldwork and AR utilization leading to the most accurate enumeration while containing costs. Staff worked on a paper and presentation containing theoretical results in decision theory as well as general operational scenarios for Census 2020. This will be presented at the 2016 Joint

Statistical Meetings in Chicago. Staff provided a draft of this paper to Decennial staff.

Staff: Darcy Steeg Morris (x33989), Yves Thibaudeau

G. Evaluation of Response Error Using Administrative Records

Description: Censuses and their evaluations ask respondents to recall where they lived on Census Day, April 1. Some interviews for evaluations take place up to eleven months after this date. Respondents are asked when they moved to their current address, and the assumption has been that respondents who move around April 1 are able to give correct answers. Error in recalling a move or a move date may cause respondents to be enumerated at the wrong location in the census. This study investigates recall error in reports of moves and move dates in censuses and sample surveys using data from survey files linked to administrative records.

Highlights: During Q3, staff continued to collaborate with staff in the Center for Survey Measurement (CSM) on analyses of recall error for reports of moves and move dates in surveys using data from survey files linked to administrative records. Staff pursued two studies. One of the studies used data from the Recall Bias Study, which was part of the 2010 Census Evaluation and Experiments Program. Results from the study were published this quarter by the journal *Survey Methods: Insights from the Field*. The other study uses data prepared for the "Memory Recall of Migration Dates in the National Longitudinal Survey of Youth" developed under a contract with the National Opinion Research Center (NORC). Staff continues to improve the draft manuscript for this study by addressing the comments received from reviewers. In addition, staff submitted an invited paper regarding lessons learned about evaluating survey data with administrative records files to the *Proceedings of the 2016 Methodology Symposium* sponsored by Statistics Canada.

Staff: Mary Mulry (x31759)

H. Special Census: Disclosure Avoidance in Group Quarters

Description: Staff works with the Decennial Information Technology Division (DITD) to create synthetic data for disclosure avoidance in group quarters data for ongoing Special Census production.

Highlights: Work on this project is now complete.

Staff: Rolando Rodriguez (x31816)

I. 2020 Unduplication Research

Description: The goal of this project is to conduct research to guide the development and assessment of methods for conducting nationwide matching and unduplication in the 2020 Decennial Census, future Censuses and other matching projects. Our staff will also develop and test new methodologies for unduplication. The project is coordinated by one of the 2020 Census Integrated Project Teams.

Highlights: Staff continued investigating networked links to identify unusual coincidental matches.

Staff: Michael Ikeda (x31756), Ned Porter, Bill Winkler, Emanuel Ben-David

1.5 ADDRESS CANVASSING IN FIELD (Decennial Project 6350B02)

A. Master Address File (MAF) Error Model and Quality Assessment

Description: The MAF is an inventory of addresses for all known living quarters in the U.S. and Puerto Rico. This project will develop a statistical model for MAF errors for housing units (HUs), group quarters (GQs), and transitory locations (TLs). This model, as well as an independent team, will be used to conduct independent quality checks on updates to the MAF and to ensure that these quality levels meet the 2020 Census requirements.

Highlights: Staff worked toward completing studies of spatial generalized mixed models on selected regions (Baltimore and New York counties) to determine the utility of accounting for spatial dependence. Using the approach of Hughes and Haran (2013) to reduce the dimension of the spatial effects, only subtle improvements are seen in predictions over nonspatial models, but regression coefficients and their significance change substantially. Spatial correlation in residuals is also reduced in spatial models. Staff considered the use of statistical decision theory to aid decisions such as "canvass" versus "do not canvass" when the underlying state of coverage error in a block takes on categories such as "Low", "Medium", and "High". Initial results emphasize the sensitivity of decisions to the decision maker's choice of utility function.

Staff: Andrew Raim (x37894), Laura Ferreira (DSSD), Krista Heim (DSSD), Scott Holan (University of Missouri)

B. Development of Block Tracking Database
Description: The Targeted Address Canvassing (TRMAC) project supports Reengineered Address Canvassing for the 2020 Census. The primary goal of the TRMAC project is to identify geographic areas to be managed in the office (i.e., in-office canvassing) and geographic areas to be canvassed in the field. The focus of the effort is on decreasing in-field and assuring the Master Address File (MAF) is current, complete, and accurate. The Block Assessment, Research, and Classification Application (BARCA) is an interactive review tool which will allow analysts to assess tabulation blocks—and later Basic Collection Units (BCUs)—by comparing housing units in 2010 imagery and current imagery, along with TIGER reference layers and MAF data.

Highlights: No significant updates this quarter.

Staff: Tom Petkunas (x33216)

C. Detection of Map Changes
Description: This research is concerned with developing statistical techniques to detect changes in maps, utilizing remote sensing data, such as LIDAR.

Highlights: During Q3 of FY 2016, staff continued meetings with Census Bureau geographers to develop methods for detecting map changes. Staff also wrote Matlab code for a modified Hough transform line detector, with results comparable to shearlet methodology.

Staff: Dan Weinberg (x38854)

1.6 AMERICAN COMMUNITY SURVEY (ACS) (Decennial Project 6385B70)

A. ACS Applications for Time Series Methods
Description: This project undertakes research and studies on applying time series methodology in support of the American Community Survey (ACS).

Highlights: During Q3 of FY 2016, staff extended methodology and R code for custom multi-year estimates to handle point estimates, as opposed to period estimates. Staff continued meetings with clients from Veterans Administration to discuss practical aspects of the project.

Staff: Tucker McElroy (x33227), Osbert Pang

B. ACS Imputation Research and Development
Description: The American Community Survey (ACS) process of editing and post-edit data-review is currently time and labor intensive. It involves repeatedly submitting an entire collection year of micro-data to an edit- enforcement program (SAS software). After each pass through the edit-enforcement program, a labor-intensive review process is conducted by a staff of analysts to identify inconsistencies and quality problems remaining in the micro-data. Before the data are ready for public release, they have at least three passes through the edit-enforcement program and three review processes by the analysts, taking upward of three months. The objective of this project is to experiment with a different strategy for editing—while keeping the same edit rules—and to assess if the new strategy can reduce the number of passes through the edit process and the duration of the review process.

Highlights: No significant updates this quarter.

Staff: Yves Thibaudeau (x31706)

C. Data Analysis of ACS CATI-CAPI Contact History
Description: This project continues earlier analyses of the American Community Survey (ACS) Computer Assisted Telephone Interview (CATI) and Computer Assisted Personal Interview (CAPI) contact history data. It focuses exclusively on CAPI with the goal of informing policy decisions on curtailing of CAPI contact attempts to minimize respondent burden on sampled households without unacceptable losses of ACS interviews.

Highlights: During Q3, staff completed revisions of an *ACS Research Memorandum* on the results of the August 2015 Pilot study of Interventions to remove cases with excessive burden from CAPI follow-up. The results of this report along with the previous ACS research reports generated in connection with this project were written in the form of a research paper submitted to *Journal of Official Statistics Special Issue on Adaptive Design.*

Staff: Eric Slud (x34991), Robert Ashmead, Todd Hughes (ACSO), Rachael Walsh (OSCA)

D. Assessing Uncertainty in ACS Ranking Tables
Description: This project presents results from applying statistical methods which provide statements of how good the rankings are in the ACS Ranking Tables (see The Ranking Project: Methodology Development and Evaluation Research Section under Projects 0331000 and 0925000).

Highlights: [See General Research: Survey Sampling-Estimation and Modeling (C), The Ranking Project: Methodology Development and Evaluation]

Staff: Tommy Wright (x31702), Martin Klein, Brett Moran, Nathan Yau

E. Confidence Intervals for Proportions in ACS Data
[See General Research: Small Area Estimation (B), Coverage Properties of Confidence Intervals for Proportions in Complex Surveys]

F. Voting Rights Section 203 Model Evaluation and Enhancements Towards Mid-Decadal Determinations
Description: Section 203 of the *Voting Rights Act* asks for determinations relating to limited English proficiency and limited education of specified small domains (race and ethnicity groups) for small areas such as counties or minor civil divisions (MCDs). Section 203 seeks to determine whether or not small areas must provide voting materials in languages other than English. Previous research undertaken provided a small area model-based estimate derived from American Community Survey (ACS) 5-year data and 2010 Census data, which provides smaller estimated variances than ACS design-based estimates in many cases. Research and groundwork into the production mid-decade determination is ongoing.

Highlights: During Q3, staff continued researching the comparison between direct survey-weighted estimates and model-based estimates to support Section 203 determinations. Model comparisons included a newly developed Dirichlet-multinomial model as well as a hierarchical Beta-Binomial model in the same spirit as the model used in 2011. Staff wrote research summaries and drafts of a technical document to aid in the decision between these models and to display the consequences of the direct estimates and the model-based estimates on determinations, based on practice analyses with ACS five-year 2012 data. A briefing document for the Associate Director for Decennial was prepared, describing the Dirichlet-Multinomial as the chosen model for 2016 Section 203 Determinations.

Staff: Patrick Joyce (x36793), Eric Slud, Robert Ashmead, Tommy Wright, Tom Louis (ADRM), John Abowd (ADRM)

1.7 DEMOGRAPHIC STATISTICAL METHODS DIVISION SPECIAL PROJECTS (Demographic Project TBA)

A. Special Project on Weighting and Estimation
Description: This project involves regular consulting with Current Population Survey (CPS) Branch staff on design, weighting, and estimation issues regarding the CPS. Issues discussed include design strategy for systematic sampling intervals, for rotating panels, composite estimation, variance estimation, and the possibility of altering CPS weighting procedures to allow for a single simultaneous stage of weight-adjustment for nonresponse and population controls.

Highlights: No significant updates this quarter.

Staff: Eric Slud (x34991), Yang Cheng (DSMD)

1.8 DEMOGRAPHIC SURVEYS DIVISION (DSD) SPECIAL PROJECTS (Demographic Project 0906/1444X00)

A. Data Integration
Description: The purpose of this research is to identify microdata records at risk of disclosure due to publicly available databases. Microdata from all Census Bureau sample surveys and censuses will be examined. Potentially linkable data files will be identified. Disclosure avoidance procedures will be developed and applied to protect any records at risk of disclosure.

Highlights: No significant updates this quarter.

Staff: Ned Porter (x31798), Marlow Lemons (CDAR)

1.9 POPULATION DIVISION PROJECTS (Demographic Project TBA)

A. Introductory Sampling Workshop
Description: In support of Population Division's International Programs Area, staff will conduct (on request) introductory sampling workshops with focus on probability sampling for participants from various countries. These workshops are primarily funded by USAID.

Highlights: No significant updates this quarter. Plans are to offer the workshop in the Fall 2016 for international participants.

Staff: Tommy Wright (x31702), Michael Leibert

1.10 SOCIAL, ECONOMIC, AND HOUSING STATISTICS DIVISION SMALL AREA ESTIMATION PROJECTS (Demographic Project 7165016)

A. Research for Small Area Income and Poverty Estimates (SAIPE)
Description: The purpose of this research is to develop, in collaboration with the Small Area Estimates Branch in the Social, Economic, and Housing Statistics Division (SEHSD), methods to produce "reliable" income and poverty estimates for small geographic areas and/or small demographic domains (e.g., poor children age 5-17 for counties). The

methods should also produce realistic measures of the accuracy of the estimates (standard errors). The investigation will include assessment of the value of various auxiliary data (from administrative records or surveys) in producing the desired estimates. Also included would be an evaluation of the techniques developed, along with documentation of the methodology.

Highlights: During Q3 of FY 2016, staff continued to explore estimating year-to-year changes in SAIPE estimates of children in poverty at the county level through a bivariate model with measurement error. Staff wrote and debugged a specialized program that can fit this model efficiently through Markov Chain Monte Carlo (MCMC). In the debugging process, several different implementations of the model were explored as well as different prior distributions for the parameters and different parametrizations. Algorithms developed for this model included an MCMC algorithm that uses a combination of Gibbs Sampling and Metropolis Hastings, as well as an importance-sampling algorithm used for comparison. The algorithms were used to compare three different models for estimating the 2010-2011 changes in county rates of school-aged children in poverty. Staff also compared results to those obtained by the software JAGS. Staff wrote a paper about this project and submitted it to a peer-reviewed journal.

Staff: Jerry Maples (x32873), Carolina Franco, William Bell (ADRM)

B. Small Area Health Insurance Estimates (SAHIE)
Description: At the request of staff from the Social, Economic, and Housing Statistics Division (SEHSD), our staff will review current methodology for making small area estimates for health insurance coverage by state and poverty level. Staff will work on selected topics of SAHIE estimation methodology, in conjunction with SEHSD.

Comparing Small Area Estimates over Time
Highlights: During Q3 of FY 2016, staff continued to develop models for inference on year-to-year change of small area proportions from different time periods using beta mixed effects regression models. Staff showed that predictions from marginal beta mixed effects models are bounded between the synthetic estimate and the direct survey estimate. Predictions tend to either the direct or synthetic estimate as either the model variance or the direct variance tends to 0, a result which coincides with predictions made using linear mixed effects models. Staff worked on developing Markov Chain Monte Carlo (MCMC) algorithms for fitting this class of models and applied the methods to both simulated data and real SAIPE and SAHIE data.

Staff: Ryan Janicki (x35725)

C. Sub County Estimates of Poverty from Multi-year ACS Data
Description: This project is from the Development Case Proposal to improve the estimates of poverty related outcomes from the American Community Survey (ACS) at the tract level. Various modeling techniques, including model-based and model-assisted, will be used to improve on the design-based multi-year estimates currently produced by the ACS. The goal is to produce more accurate estimates of poverty and income at the tract level and develop a model framework that can be extended to outcomes beyond poverty and income.

Highlights: During Q3 of FY 2016, staff investigated models for the number of poor school-aged children in tracts using three different variants of overdispersed Poisson models: CMP (Conway and Maxwell), Generalized Poisson (Consul and Jian) and Double Exponential family for Poisson (Efron), in addition to the negative Binomial distribution. The different distributions will be tested against the artificial population data samples. These overdispersed count distributions will then be used to form an area-level small area model to predict the number of school-age children in poverty.

Staff: Jerry Maples (x32873), Ryan Janicki, Carolina Franco, William Bell (ADRM)

1.11 ECONOMIC STATISTICAL COLLECTION (Economic Project 1183X01)

1.12 ECONOMIC MONTHLY/RETAIL (Economic Project 1001X00)

A. Research on Imputation Methodology for the Monthly Wholesale Trade Survey
Description: In the previous phase of this project, staff conducted a simulation study to investigate new imputation methodology for the Monthly Wholesale Trade Survey (MWTS). In this phase of the project, staff are creating a more realistic simulated wholesale trade population and investigating improved MWTS estimators. The MWTS is a longitudinal survey that provides month-to-month information on sales and inventories of U.S. merchant wholesalers. Key estimates produced from this survey include total sales, month-to-month relative change in sales, total inventories, and month-to-month relative change in inventories (overall and within industry subclasses). There are a number of challenges when developing estimators for the MWTS, including variables with highly skewed distributions, missing values in predictor variables from the Economic Census, and survey variables with trends that differ across industry classes.

The longitudinal information in addition to a rich set of frame data available from the Economic Census can be used to build Bayesian models that address these challenges. It is expected that this model will be applicable to other business surveys.

Highlights: Staff continued developing a new version of the realistic artificial population used for drawing simulated samples of MWTS data. The constructed population enables us to study the statistical properties of new estimation and imputation procedures on the MWTS. Staff are developing the population in two parts: (1) units that are sampled with certainty, and (2) remaining units which are sampled based on a stratified random sampling design. To construct the certainty population, we used observed MWTS data over a two-year period to obtain a roster of sampling units, reporting units, and tabulation units. These units were then merged with data from the sampling frame and with data from the Economic Census. Missing values in the certainty population were then imputed using multivariate imputation by chained equations with random forests as the conditional models. Staff experimented with variations of this imputation method and analyzed the outputs in order to find a good imputation model. To construct the population of non-certainty units, staff used data from the sampling frame to obtain a roster of units and merged it with MWTS and Economic Census data. The next step is to fill in any missing data to get a complete non-certainty population. As with the certainty population, staff are using multivariate imputation by chained equations with random forests as the conditional models to impute missing data in the non-certainty population. Because many non-certainty units are not in the MWTS sample, a very large proportion of the non-certainty population have all values missing for the MWTS variables. This large-scale imputation poses some challenges. To avoid having similar values repeated many times in the population, staff modified the random forest imputation procedure so that it draws the imputed value from a kernel density fitted to the candidate values. Work on imputation of the non-certainty population is ongoing.

Staff: Martin Klein (x37856), Joe Schafer (ADRM), Joanna Fane Lineback (ADEP), Brett Moran

B. Use of Big Data for Retail Sales

Description: In this project, we are investigating the use of "Big Data" to fill gaps in retail sales estimates currently produced by the Census Bureau. Specifically, we are interested in how to use "Big Data" to supplement existing monthly/annual retail surveys with a primary focus on exploring (1) how to use third party data to produce geographic level estimates more frequently than once every five years (i.e. a new product), and (2) the possibility of using third party data tabulations to improve/enhance Census Bureau estimates of monthly retail sales - for example, validation and calibration. Various types of data are being pursued such as credit card transaction data and scanner data.

Highlights: During Q3 of FY 2016, staff continued working with Economic Directorate staff to assess the quality and potential use of aggregated electronic transaction data from First Data. Staff participated in working meetings with staff from Palantir, the company that houses, manages and visualizes the First Data data. Staff provided general input and suggestions for the visualization and computing tool that Palantir developed for analysis of the First Data data. Staff studied small area estimation models for assessing the predictive power of the First Data transaction data for industry/state level estimates of monthly change and retail sales totals. Staff contributed to a summary report of this research and reviewed a proposal for future work.

Staff: Darcy Steeg Morris (x33989), Osbert Pang, Tommy Wright, Tucker McElroy, Brian Monsell, Bill Bostic (ADEP), Scott Scheleur (SSSD), Bill Davie, Jr. (ESMD)

1.13 ECONOMIC CENSUS/SURVEY ENGINEERING: TIME SERIES RESEARCH; ECONOMIC MISSING DATA/PRODUCT LINE DATA; DEVELOPMENT/SAS (Economic Project 2220B10)

A. Seasonal Adjustment Support

Description: This is an amalgamation of projects whose composition varies from year to year but always includes maintenance of the seasonal adjustment software used by the Economic Directorate.

Highlights: Staff provided seasonal adjustment and software support for users within and outside the Census Bureau including Palentir, EJJE (Mexico), Reserve Bank of New Zealand, Bureau of Economic Analysis, Bureau of Labor Statistics, Statistics New Zealand, Office of National Statistics (UK), Statistics Centre of Abu Dhabi, Barcelona Graduate School for Economics, University of Bern.

Staff met with Ronald Indergand of the University of Bern to discuss his work comparing seasonal adjustment revisions from X-11 and SEATS. Staff worked with other Census staff to answer questions for a *Wall Street Journal* reporter doing an article on seasonal adjustment.

Staff organized a *2016 Summer at Census* visit by Christophe Sax, a consultant who has developed the seasonal R package that interfaces with X-13ARIMA-SEATS. Sax gave a talk and had discussions with

Census Bureau personnel on future work related to the seasonal package.

Staff participated in a briefing on seasonal adjustment practice at the Census Bureau to Mr. Silvio Peruzzo of Rokos Capital Management LLP on behalf of the Italian government.

Staff met with staff from CES on a plan to produce seasonal adjustments of various quarterly series (both state and national level) that are not currently being seasonally adjusted. Staff met with Economic Directorate staff to discuss residual seasonality in the Value Put in Place series. Staff continued working with Economic Statistical Methods Division (ESMD) staff to plan a seasonal adjustment workshop to be held in November 2016.

Staff: Brian Monsell (x31721), Tucker McElroy, James Livsey, Osbert Pang, Anindya Roy, Thomas Trimbur, William R. Bell (ADRM)

B. Seasonal Adjustment Software Development and Evaluation

Description: The goal of this project is a multi-platform computer program for seasonal adjustment, trend estimation, and calendar effect estimation that goes beyond the adjustment capabilities of the Census X-11 and Statistics Canada X-11-ARIMA programs, and provides more effective diagnostics. The goals for FY 2015 include: continuing to develop a version of the X-13ARIMA-SEATS program with accessible output and updated source code so that, when appropriate, the Economic Directorate can produce SEATS adjustments; and incorporating further improvements to the X-13ARIMA-SEATS user interface, output and documentation. In coordination and collaboration with the Time Series and Related Methods Staff of the Economic Statistical Methods Division (ESMD), staff will provide internal and/or external training in the use of X-13ARIMA-SEATS and the associated programs, such as X-13-Graph, when appropriate.

Highlights: Staff revised the code related to user-defined regressors to add user-defined types for trading day, constant, length of month/quarter, and outliers (AO, LS and SO), and checked how multiple types of user-defined regressors worked with the chi-squared and F-statistics. Staff also fixed defects in the irregular regression modeling routines and the SEATS routines. Staff added a quarterly seasonality test for monthly time series, an option to the spectrum spec to generate quarterly seasonality diagnostics for monthly series, timer information to the diagnostics files with the specification of a runtime argument (-t), and an easter[0] regressor to the regARIMA model to model lower levels of economic activity for the day of the Easter holiday.

Staff is currently developing a library of the X-13ARIMA-SEATS routines, and continues to develop a version of the X-13ARIMA-SEATS program with updated SEATS routines.

Staff continued the development of sigex, a suite of R routines for modeling multivariate time series. Staff revised the software to include a windowing method to perform the signal extraction, and improve modeling procedures. Staff compiled the source code for the latest version of regCMPNT, fixed defects in the code, and ran examples to test the executable.

Staff: Brian Monsell (x31721), Tucker McElroy, Osbert Pang

C. Research on Seasonal Time Series - Modeling and Adjustment Issues

Description: The main goal of this research is to discover new ways in which time series models can be used to improve seasonal and calendar effect adjustments. An important secondary goal is the development or improvement of modeling and adjustment diagnostics. This fiscal year's projects include: (1) continuing research on goodness of fit diagnostics (including signal extraction diagnostics and Ljung-Box statistics) to better assess time series models used in seasonal adjustment; (2) studying the effects of model based seasonal adjustment filters; (3) studying multiple testing problems arising from applying several statistics at once; (4) determining if information from the direct seasonally adjusted series of a composite seasonal adjustment can be used to modify the components of an indirect seasonal adjustment, and more generally investigating the topics of benchmarking and reconciliation for multiple time series; (5) studying alternative models of seasonality, such as Bayesian and/or long memory models and/or heteroskedastic models, to determine if improvement to seasonal adjustment methodology can be obtained; (6) studying the modeling of stock holiday and trading day on Census Bureau time series; (7) studying methods of seasonal adjustment when the data are no longer univariate or discrete (e.g., multiple frequencies or multiple series); (8) studying alternative seasonal adjustment methods that may reduce revisions or have alternative properties; and (9) studying nonparametric methods for estimating regression effects, and their behavior under long range dependence and/or extreme values.

Highlights: Staff (a) continued simulation and empirical work to assess new methodologies for fitting vector moving average models; (b) developed algorithms for quickly computing signal extraction estimates from long samples of high-frequency data; (c) studied the generation of meteorological regressors based on weather data from the National Climatic Data Center (obtained through a web scraping tool), to be used in a

weather-assisted seasonal adjustment of construction series; (d) continued modeling of daily time series (New Zealand immigration data and credit card transaction data), with multiple periods of seasonality, and utilized software from the Bureau of Labor Statistics to obtain seasonal adjustments; and (e) examined mean squared error in simulation of seasonal adjustment from the X11 and SEATS software.

Staff: Tucker McElroy (x33227), James Livsey, Brian Monsell, Osbert Pang, William Bell (ADRM), David Findley (Private Collaborator)

D. Supporting Documentation and Software for X-13ARIMA-SEATS

Description: The purpose of this project is to develop supplementary documentation and utilities for X-13ARIMA-SEATS that enable both inexperienced seasonal adjustors and experts to use the program as effectively as their backgrounds permit. This fiscal year's goals include improving the X-13ARIMA-SEATS documentation, exploring the use of R packages that interface with X-13ARIMA-SEATS, and exploring the use of component and Java software developed at the National Bank of Belgium.

Highlights: Staff updated the *X-13ARIMA-SEATS REFERENCE MANUAL* to include information on new options and diagnostics added to working versions of the software. Staff continued working with Christoph Sax on development of R package, x13story, to facilitate communication of X-13ARIMA-SEATS related topics.

Staff began working with CSRM and CSM staff to facilitate the reorganization of the research report series. Staff developed a generalized routine to develop moving holiday regressors for weekly and daily time series. Staff developed a modified Hough transform line detector in Matlab.

Staff: Brian Monsell (x31721), Tucker McElroy, James Livsey, Osbert Pang, Daniel Wienberg

E. Missing Data Adjustment Methods for Product Data in the Economic Census

Description: The Economic Census collects general items from business establishments such as total receipts, as well as more detailed items such as product sales. Although product data are an essential component of the Economic Census, item response rate is low. This project investigates methods for imputing missing product data in the Economic Census. Staff researched three methods for treating missing product line data: expansion estimation, hot deck (random and nearest neighbor), and sequential regression multivariate imputation (SRMI). Staff was asked to apply the SRMI method to these data and assist in making a recommendation.

Highlights: Work on this project is complete.

Staff: Darcy Steeg Morris (x33989), Maria Garcia, Yves Thibaudeau

1.14 2012 COMMODITY FLOW SURVEY (Economic Project 7103012)

A. 2012 Commodity Flow Survey

Description: This project provides a retrospective analysis of the cost-quality tradeoffs that the Commodity Flow Survey (CFS) made moving from a 2007 paper-only to a 2012 paper and electronic multi-mode data collection strategy. Based on the data quality findings, the possibility of adding additional edits or modifications to the instruments will be investigated. Optimization strategies for a multi-mode data collection strategy in the 2017 CFS and cost-quality implications of an all-electronic collection will be studied.

Highlights: No significant updates this quarter.

Staff: Robert Ashmead (x31564), Eric Slud, Joanna Fane Lineback (CSM)

1.15 INVESTIGATION OF ALTERNATIVE METHODS FOR RESOLVING BALANCE COMPLEX FAILURES IN StEPS (Economic Project TBA)

A. Investigation of Alternative Methods for Resolving Balance Complex Failures in StEPS

Description: The Standard Economic Processing System (StEPS) implements a raking algorithm for adjusting balance complexes in order to satisfy the requirement that the sum of items (details) in a balance complex balances to reported totals. In this project, we research alternative methods to raking when the data items are negative or when there is subtraction in the balance complex.

Highlights: The StEPS generalized system implements a raking algorithm for adjusting failing balance complexes. The raking adjustment fails when the balance complex includes subtraction or if detail items are allowed to take negative values. During Q3, staff developed four separate alternatives methods to handle these cases. All four methods rely on solving a linear programming problem; the objective functions minimize the change between the reported details while the constraints ensure the raked details add up to the

total. Staff met with subject matter analysts to discuss how they manually resolved most failing balance complexes. We plan to incorporate their techniques when implementing the proposed methods.

Staff: Maria Garcia (x31703), Yves Thibaudeau

1.16 BUSINESS DYNAMICS STATISTICS—EXPORT FILE WEIGHTING ISSUE (Research and Methodology Directorate TBA)

A. Business Dynamics Statistics—Export File Weighting Issue

Description: The challenge: we are unable to match the universe of export transactions to firms on the business register. Therefore, we cannot identify the universe of firms that export U.S. goods. We can pursue two options—(i) produce business dynamics statistics based on the identified cases only. For example, an official Census Bureau data product, the *Profile of U.S. Importing and Exporting Companies*, is released based on the "known" matches and users are provided with a technical documentation explaining the data limitations; or (ii) construct weights to create business dynamics statistics that are representative of the U.S. exporter population.

Highlights: During Q3, staff researched the possibility of constructing weights using data from other Census Bureau data sources (e.g. Economic Census Data). In discussions with Foreign Trade and Economic Census data experts, staff concluded this is not a valid solution and proposed two alternative approaches. One such approach is to calculate a measure of the quality of the existing matches and their representativeness in the exports and patents data files using the R-indicator introduced in Schouten, Cobben and Bethlehem (2009). Alternatively, staff proposed setting up the problem of augmenting the exports and patents datasets with variables from the Business Register (BR) as a missing data problem and implementing two separate imputation methods: Statistical Matching (D'Orazio, 2016) and the multiple imputation procedure Sequential Regression Multivariate Imputation (Raghunathan et al., 2001).

Staff: Maria Garcia (x31703), Emanuel Ben-David

1.17 NATIONAL SURVEY OF DRUG USE & HEALTH (Census Bureau Project 7236045)

A. National Survey of Drug Use & Health

Description: This project is a feasibility study concerning the extension of the National Survey of Drug Use & Health (NSDUH) to Puerto Rico and other U.S. island areas. Our staff will focus specifically on small area estimation methodology and will determine if and how the island areas can be incorporated into the current NSDUH small area estimation methodology.

Highlights: No significant updates this quarter.

Staff: Robert Ashmead (x31564), Jerry Maples

1.18 PROGRAM DIVISION OVERHEAD (Census Bureau Project 0331000)

A. Center Leadership and Support

This staff provides ongoing leadership and support for the overall collaborative consulting, research, and administrative operation of the center.

Staff: Tommy Wright (x31702), Lauren Emanuel, Michael Hawkins, Michael Leibert, Erica Magruder, Eric Slud, Kelly Taylor

B. Research Computing

Description: This ongoing project is devoted to ensuring that Census Bureau researchers have the computers and software tools they need to develop new statistical methods and analyze Census Bureau data.

Highlights: During Q3, the Integrated Research Environment (IRE) team continued to develop the IRE, a shared Linux computing platform that will replace the current "compute clusters" research1, research2, and the RDC cluster. The IRE will provide the logical separation of project data and activities currently provided in the RDC environment, but without using a separate login for each project. A collection of scripts will enable the user to "change into" a particular project where they will be presented only with the data associated with that project. Testing of those scripts and integrating them with the job scheduler (PBSPro) is the current focus. Progress was made in setting the correct DISPLAY variable in interactive PBS sessions to enable the forwarding of X output. It is still necessary to set the DISPLAY in the shell that launches the interactive PBS session. A wrapper script for qsub is being considered as a solution for interactive sessions. Further testing is needed to see how various statistical software behaves when run

within the modified shell presented to it by the system. Software that spawns "worker" processes either on the same node or on other nodes (e.g., Matlab's Parallel Computing Toolbox) needs to be tested to ensure that the worker processes have the appropriate view of the filesystem namespace and operate as expected.

Staff: Chad Russell (x33215)

2. RESEARCH

2.1 GENERAL RESEARCH AND SUPPORT (Census Bureau Project 0331000)

2.2 GENERAL RESEARCH (Census Bureau Project 0925000)

Missing Data, Edit, and Imputation

Motivation: Missing data problems are endemic to the conduct of statistical experiments and data collection projects. The instigators almost never observe all the outcomes they had set to record. When dealing with sample surveys or censuses that means individuals or entities in the survey omit to respond, or give only part of the information they are being asked to provide. In addition the information provided may be logically inconsistent, which is tantamount to missing. To compute official statistics, agencies need to compensate for missing data. Available techniques for compensation include cell adjustments, imputation and editing. All these techniques involve mathematical modeling along with subject matter experience.

Research Problems: Compensating for missing data typically involves explicit or implicit modeling. Explicit methods include Bayesian multiple imputation and propensity score matching. Implicit methods revolve around donor-based techniques such as hot-deck imputation and predictive mean matching. All these techniques are subject to edit rules to ensure the logical consistency of remedial product. Research on integrating together statistical validity and logical requirements into the process of imputing continues to be challenging. Another important problem is that of correctly quantifying the reliability of predictors that have been produced in part through imputation, as their variance can be substantially greater than that computed nominally.

Potential Applications: Research on missing data leads to improved overall data quality and predictors accuracy for any census or sample survey with a substantial frequency of missing data. It also leads to methods to adjust the variance to reflect the additional uncertainty created by the missing data. Given the ever rising cost of conducting censuses and sample surveys, imputation and other missing-data compensation methods may come to replace actual data collection, in the future, in situations where collection is prohibitively expensive.

A. Editing

Description: This project covers development of methods for statistical data editing. Good methods allow us to produce efficient and accurate estimates and higher quality microdata for analyses.

Highlights: No significant updates this quarter.

Staff: Maria Garcia (x31703)

B. Editing and Imputation

Description: Under this project, our staff provides advice, develops computer edit/imputation systems in support of demographic and economic projects, implements prototype production systems, and investigates edit/imputation methods.

Highlights: Staff researched the possibility of increasing the accuracy of the predictors on mortgage ownership derived from the American Community Survey (ACS) by extracting and exploiting information from the "CoreLogic" commercial "Big Data" data set. Staff is exploring the possibility of creating "equal-propensity strata" where the value of the mortgage question is homogeneous. This could provide predictive information to improve the accuracy of the mortgage item in ACS.

Staff: Yves Thibaudeau (x31706), Maria Garcia, Martin Klein, Darcy Steeg Morris, Bill Winkler

Record Linkage

Motivation: Record linkage is intrinsic to efficient, modern survey operations. It is used for unduplicating and updating name and address lists. It is used for applications such as matching and inserting addresses for geocoding, coverage measurement, Primary Selection Algorithm during decennial processing, Business Register unduplication and updating, re-identification experiments verifying the confidentiality of public-use microdata files, and new applications with groups of administrative lists. Significant theoretical and algorithmic progress (Winkler 2004ab, 2006ab, 2008, 2009a; Yancey 2005, 2006, 2007, 2011) demonstrates the potential for this research. For cleaning up administrative records files that need to be linked, theoretical and extreme computational results (Winkler 2010, 2011b) yield methods for editing, missing data and even producing synthetic data with valid analytic properties and reduced/eliminated re-identification risk. Easy means of constructing synthetic make it straightforward to pass files among groups.

Research Problems: The research problems are in three major categories. First, we need to develop effective ways of further automating our major record linkage operations. The software needs improvements for matching large sets of files with hundreds of millions of records against other large sets of files. Second, a key open research question is how to effectively and automatically estimate matching error rates. Third, we need to investigate how to develop effective statistical analysis tools for analyzing data from groups of administrative records when unique identifiers are not available. These methods need to show how to do correct demographic, economic, and statistical analyses in the presence of matching error.

Potential Applications: Presently, the Census Bureau is contemplating or working on many projects involving record linkage. The projects encompass the Demographic, Economic, and Decennial areas.

A. Disclosure Avoidance for Microdata
Description: Our staff investigates methods of microdata masking that preserves analytic properties of public-use microdata and avoid disclosure.

Highlights: Staff refereed two papers for *Statistical Data Protection 2016*. Staff met with a professor from Cornell Tech to discuss issues related to analytic properties of anonymized data. Staff provided a set of papers (Winkler 1997, 2003, 2008, 2010) about analytic and probabilistic constraints on data that can be used to reduce re-identification risk in public-use microdata. Staff met with one individual in the Center for Disclosure Avoidance (CDAR) to discuss microdata confidentiality issues and computational algorithms. Staff met with the Research and Methodology Associate Director to discuss issues related to microdata confidentiality, particularly theoretical extensions of differential privacy to loglinear models.

Staff: William Winkler (x34729)

B. Record Linkage and Analytic Uses of Administrative Lists
Description: Under this project, staff will provide advice, develop computer matching systems, and develop and perform analytic methods for adjusting statistical analyses for computer matching error.

Highlights: Staff began comparisons of existing theoretical methodologies for regression analysis of linked data. The main goal of this comparison is to take an approach that can improve upon these methods. Staff e-mailed considerable information about record linkage to a staff member at the National Agriculture Statistical Service who is working on the main list development for the 2017 Agriculture Census. CSRM staff, along with Economic Directorate programming staff, created the list-development systems for the 1992 and 1997 Agriculture Censuses.

Staff e-mailed considerable information on specific record linkage methods to a staff member at the National Cancer Institute. Staff agreed to be on a Ph.D. committee at the University of Maryland. Staff also e-mailed considerable information related to record linkage to a member of the Federal Reserve Board.

Staff: William Winkler (x34729), Ned Porter, Emanuel Ben-David

C. Modeling, Analysis, and Quality of Data
Description: Our staff investigates methods of the quality of microdata primarily via modeling methods and new software techniques that accurately describe one or two of the analytic properties of the microdata.

Highlights: Staff provided considerable advice, papers, and software to professors and students at Carnegie-Mellon University who are working on record linkage projects funded by the National Science Foundation and by the Decennial Statistical Studies Division. Staff provided a large draft chapter on quality methods for administrative lists that included basic record linkage, modeling/edit/imputation, and recent methods for adjusting statistical analyses for linkage error. The final draft of the chapter "Record Linkage" by Professor Peter Christen and one staff member was accepted in the *Encyclopedia of Machine Learning and Data Mining*.

Staff worked on extending the algorithms for generating implicit edits. The extension uses heuristics that keep track of the most active fields used in generating new edits.

Staff: William Winkler (x34729), Ned Porter, Emanuel Ben-David, Maria Garcia

D. R Users Group
Description: The initial objective of the R Users Group is to identify the areas of the Census Bureau where R software is developed and those other areas that could benefit from such development. The scope of the topics is broad and it includes estimation, missing data methods, statistical modeling, Monte-Carlo and resampling methods. The ultimate goal is to move toward integrated R tools for statistical functionality at the Census Bureau.

Initially the group will review basic skills in R and provide remedial instruction as needed. The first topic for deeper investigation is complex-survey infrastructure utilities, in particular an evaluation of the "Survey" package and its relevance at the Census Bureau in the context of weighing, replication, variance estimation and other structural issues.

Highlights: No significant updates this quarter.

Staff: Yves Thibaudeau (x31706), Chad Russell

Small Area Estimation

Motivation: Small area estimation is important in light of a continual demand by data users for finer geographic detail of published statistics. Traditional demographic surveys designed for national estimates do not provide large enough samples to produce reliable direct estimates for small areas such as counties and even most states. The use of valid statistical models can provide small area estimates with greater precision, however bias due to an incorrect model or failure to account for informative sampling can result. Methods will be investigated to provide estimates for geographic areas or subpopulations when sample sizes from these domains are inadequate.

Research Problems:
• Development/evaluation of multilevel random effects models for capture/recapture models.
• Development of small area models to assess bias in synthetic estimates.
• Development of expertise using nonparametric modeling methods as an adjunct to small area estimation models.
• Development/evaluation of Bayesian methods to combine multiple models.
• Development of models to improve design-based sampling variance estimates.
• Extension of current univariate small-area models to handle multivariate outcomes.

Potential Applications:
• Development/evaluation of binary, random effects models for small area estimation, in the presence of informative sampling, cuts across many small area issues at the Census Bureau.
• Using nonparametric techniques may help determine fixed effects and ascertain distributional form for random effects.
• Improving the estimated design-based sampling variance estimates leads to better small area models which assumes these sampling error variances are known.
• For practical reasons, separate models are often developed for counties, states, etc. There is a need to coordinate the resulting estimates so smaller levels sum up to larger ones in a way that correctly accounts for accuracy.
• Extension of small area models to estimators of design- base variance.

A. Small Area Methods with Misspecification
Description: In this project, we undertake research on area-level methods with misspecified models, primarily directed at development of diagnostics for misspecification using robust sandwich-formula variances, cross-validation, and others, and on Bayesian estimation of model parameters within two-component Fay-Herriot models.

Highlights: This project is on hold until additional staffing is available.

Staff: Jerry Maples (x32873), Gauri Datta, Eric Slud

B. Coverage Properties of Confidence Intervals for Proportions in Complex Surveys
Description: This is primarily a simulation project to investigate the coverage behavior of confidence intervals for proportions estimated in complex surveys. The goal is ultimately to inform recommendations for interval estimates in the American Community Survey (ACS), so the issues of main interest are:
(i) whether the current Wald-type intervals (defined as a point-estimator plus or minus a margin-or-error (MOE) estimate) can be improved by empirical-Bayes modifications or by modified forms of intervals known to perform well in the setting of binomial proportion-estimators, (ii) whether failures of coverage in a simulated complex survey can be ascribed to poor estimation of effective sample size or to other aspects of inhomogeneity and clustering in proportions within realistically complex populations, and (iii) whether particular problems arising with coverage of intervals for small proportions can be overcome. Future research might address whether the confidence interval methods developed for single-domain design-based estimates can also be adapted to small area estimates that borrow strength across domains.

Highlights: During Q3 of FY 2016, staff incorporated an additional way to estimate the sampling variance and to estimate the design effect into the simulation software. Staff began evaluating the new methods and debugging the software. Staff also explored the conjecture that the failure of methods to account for the uncertainty in estimation of the design effect is a main reason for tendency of undercoverage for confidence intervals for proportions in complex surveys. Staff found empirical evidence supporting this conjecture through the simulation study, and devised ideas on how to exploit this finding.

Staff: Carolina Franco (x39959), Eric Slud, Thomas Louis (ADRM), Rod Little (University of Michigan)

C. Small Area Estimates of Disability

Description: This project is from the Development Case proposal to create subnational estimates of specific disability characteristics (e.g., number of people with autism). This detailed data is collected in a supplement of the Survey of Income and Program Participation (SIPP). However, the SIPP is only designed for national level estimates. This project is to explore small area models to combine SIPP with the large sample size of the American Community Survey to produce state and county level estimates of reasonable quality.

Highlights: During Q3 of FY 2016, staff waited to hear back about a paper submitted to the *Journal of the Royal Statistical Society: Series A.*

Staff: Jerry Maples (x32873), Amy Steinweg (SEHSD)

D. Using ACS Estimates to Improve Estimates from Smaller Surveys via Bivariate Small Area Estimation Models

Description: Staff will investigate the use of bivariate area-level models to improve small area estimates from one survey by borrowing strength from related estimates from a larger survey. In particular, staff will explore the potential of borrowing strength from estimates from the American Community Survey, the largest U.S. household survey, to improve estimates from smaller U.S. surveys, such as the National Health Interview Survey, the Survey of Income and Program Participation, and the Current Population Survey.

Highlights: No significant updates this quarter.

Staff: Carolina Franco (x39959), William R. Bell (ADRM)

E. Multivariate Fay-Harriot Hierarchical Bayesian Estimation of Small Area Means under Functional Measurement Error

Description: Area-level models have been extensively used in small area estimation to produce model-based estimates of a population characteristic for small areas (e.g., Fay and Herriot, 1979). Multivariate area level models have also been used to jointly model multiple characteristics of correlated responses (e.g., Huang and Bell, 2012, Franco and Bell, 2015). Such models may lead to more precise small area estimates than separate univariate modeling of each characteristic. Typically both univariate and multivariate small area estimation models use auxiliary information to borrow strength from other areas and covariates associated with a response variable or a response vector. However, auxiliary variables are sometimes measured or obtained from sample surveys and are subject to measurement or sampling error. Researchers recognized that ignoring measurement error in the covariates and using standard solutions developed for covariates measured without error may lead to suboptimal inference. It was demonstrated in the univariate small area estimation setup that this naïve approach can result in model-based small area estimators that are more variable than the direct estimators when some of the covariate values in a small area are measured with substantial error (cf. Ybarra and Lohr, 2008, *Biometrika*; Arima, Datta and Liseo, 2015, *Scandinavian Journal of Statistics*). We are investigating a multivariate Fay-Herriot model and develop Bayes small area estimates when one or more auxiliary variables are measured with error. We work out a hierarchical Bayesian analysis for the multivariate Fay-Herriot model with a functional measurement error treatment for the covariates measured with error.

Highlights: During Q3 of FY 2016, staff developed an efficient R program that can fit Hierarchical Bayes Multivariate Fay-Herriot Models with Measurement error of dimension k with the proposed class of prior distributions. Staff discovered that plain Gibbs sampling is inefficient for this type of model and other similar linear models. In fact, Gibbs sampling is so slow that it is not feasible to apply it to large data sets (i.e., county-level school-aged children in poverty) even for relatively low dimensions (i.e., bivariate models). Staff developed an algorithm that uses a combination of Metropolis-Hastings and Gibbs sampling and can be applied to models of high dimensions as well as to large data sets. Staff did extensive debugging and compared the results to those obtained via JAGS and through an importance-sampling algorithm. Staff also applied the algorithms to SAIPE data. Staff collected the empirical, theoretical, and computational results and completed a paper that was submitted to a peer-reviewed journal.

Staff: Carolina Franco (x39959), Gauri Datta, William R. Bell (ADRM)

F. Smoothing Design Effects for Small Sample Areas

Description: In Small Area Estimation, the design-based estimates for many areas are based on very small samples. We propose using information from a larger aggregate, whose design-based variance estimator can be reliably estimated to inform us about the design effect at the small component area. Our goal is to create a principled method to use information about design effects at the higher level to estimate design effects at the lower level. Due to the lack of data, this will require strong assumptions and large amounts of smoothing of design features over the small local areas.

Highlights: During Q3 of FY 2016, staff developed methods to estimate average residual design effect over a collection of areas after accounting for unequal sampling probabilities and intraclass correlation due to clustering and informative sampling. In applications where these design effects are used to smooth out the design-based sampling variance estimates, procedures were developed when there was a model for the

underlying rate (with associated covariate predictors) and when there was no underlying model (no covariate predictors). When there were no predictors, an optimization criterion was used to determine the best weighted average of the local area level estimated rate and the group-level estimated rate.

Staff: Jerry Maples (x32873)

Survey Sampling-Estimation and Modeling

Motivation: The demographic sample surveys of the Census Bureau cover a wide range of topics but use similar statistical methods to calculate estimation weights. It is desirable to carry out a continuing program of research to improve the accuracy and efficiency of the estimates of characteristics of persons and households. Among the methods of interest are sample designs, adjustments for non-response, proper use of population estimates as weighting controls, small area estimation, and the effects of imputation on variances.

The Economic Directorate of the Census Bureau encounters a number of issues in sampling and estimation in which changes might increase the accuracy or efficiency of the survey estimates. These include, but are not restricted to, a) estimates of low-valued exports and imports not currently reported, b) influential values in retail trade survey, and c) surveys of government employment.

The Decennial Census is such a massive undertaking that careful planning requires testing proposed methodologies to achieve the best practical design possible. Also, the U.S. Census occurs only every ten years and is the optimal opportunity to conduct evaluations and experiments with methodologies that might improve the next census. Sampling and estimation are necessary components of the census testing, evaluations, and experiments. The scale and variety of census operations require an ongoing research program to achieve improvements in methodologies. Among the methods of interest are coverage measurement sampling and estimation, coverage measurement evaluation, evaluation of census operations, uses of administrative records in census operations, improvements in census processing, and analyses that aid in increasing census response.

Research Problems:

• How can methods making additional use of administrative records, such as model-assisted and balanced sampling, be used to increase the efficiency of household surveys?

• Can non-traditional design methods such as adaptive sampling be used to improve estimation for rare characteristics and populations?

• How can time series and spatial methods be used to improve ACS estimates or explain patterns in the data?

• Can generalized weighting methods be implemented via optimization procedures that allow better understanding of how the various steps relate to each other?

• Some unusual outlying responses in the surveys of retail trade and government employment are confirmed to be accurate, but can have an undesired large effect on the estimates - especially estimates of change. Procedures for detecting and addressing these influential values are being extended and examined through simulation to measure their effect on the estimates, and to determine how any such adjustment best conforms with the overall system of estimation (monthly and annual) and benchmarking.

• What models aid in assessing the combined effect of all the sources of estimable sampling and nonsampling error on the estimates of population size?

• How can administrative records improve census coverage measurement, and how can census coverage measurement data improve applications of administrative records?

• What analyses will inform the development of census communications to encourage census response?

• How should a national computer matching system for the Decennial Census be designed in order to find the best balance between the conflicting goals of maximizing the detection of true duplicates and minimizing coincidental matches? How does the balance between these goals shift when modifying the system for use in other applications?

• What can we say about the additional information that could have been obtained if deleted census persons and housing units had been part of the Census Coverage Measurement (CCM) Survey?

Potential Applications:

• Improve estimates and reduce costs for household surveys via the introduction of additional design and estimation procedures.

• Produce improved ACS small area estimates through the use of time series and spatial methods.

• Apply the same weighting software to various surveys.

• New procedures for identifying and addressing influential values in the monthly trade surveys could provide statistical support for making changes to weights or reported values that produce more accurate estimates of month-to-month change and monthly level. The same is true for influential values in surveys of government employment.

• Provide a synthesis of the effect of nonsampling errors on estimates of net census coverage error, erroneous enumerations, and omissions and identify the types of nonsampling errors that have the greatest effects.

• Describe the uncertainty in estimates of foreign-born immigration based on American Community

Survey (ACS) used by Demographic Analysis (DA) and the Postcensal Estimates Program (PEP) to form estimates of population size.
• Improve the estimates of census coverage error.
• Improve the mail response rate in censuses and thereby reduce the cost.
• Help reduce census errors by aiding in the detection and removal of census duplicates.
• Provide information useful for the evaluation of census quality.
• Provide a computer matching system that can be used with appropriate modifications for both the Decennial Census and several Decennial-related evaluations.

A. Household Survey Design and Estimation
[See Demographic Projects]

B. Sampling and Estimation Methodology: Economic Surveys
Description: The Economic Directorate of the Census Bureau encounters a number of issues in sampling and estimation in which changes might increase the accuracy or efficiency of the survey estimates. These include estimates of low-valued exports not currently reported, alternative estimation for the *Quarterly Financial Report*, and procedures to address nonresponse and reduce respondent burden in the surveys. Further, general simulation software might be created and structured to eliminate various individual research efforts. An observation is considered influential if the estimate of total monthly revenue is dominated by its weighted contribution. The goal of the research is to find methodology that uses the observation but in a manner that assures its contribution does not dominate the estimated total or the estimates of period-to-period change.

Highlights: During Q3, staff continued collaborating with staff in the Economic Directorate on research to find a statistical procedure for detecting and treating verified influential values in economic surveys to replace the current subjective procedure performed by analysts. Recent research has focused on finding an automated procedure with the expectation that any adjustments be reviewed. Previous research identified an M-estimation methodology as the most suitable choice, but the initial parameter settings for the M-estimation algorithm affect its performance. Using historical data from the Monthly Wholesale Trade Survey (MWTS), staff developed an automated data-driven approach for determining the initial parameter settings for the M-estimation algorithm parameters. A presentation on the methodology at the Fifth International Conference on Establishment Surveys received positive comments from experts in the field. Plans are to conduct a side-by-side test conducted MWTS data collection to make further assessments and possible refinements to the methodology.

Staff: Mary Mulry (x31759)

C. The Ranking Project: Methodology Development and Evaluation
Description: This project undertakes research into the development and evaluation of statistical procedures for using sample survey data to rank several populations with respect to a characteristic of interest. The research includes an investigation of methods for quantifying and presenting the uncertainty in an estimated ranking of populations. As an example, a series of ranking tables are released from the American Community Survey in which the fifty states and the District of Columbia are ordered based on estimates of certain characteristics of interest.

Highlights: Staff continued development of a draft website; obtained ten years of ACS data on more than 85 variables for ranking; and continued to experiment with various visualizations.

Staff: Tommy Wright (x31702), Martin Klein, Jerzy Wieczorek (Carnegie Mellon University), Brett Moran, Nathan Yau, Michael Leibert

D. Sampling and Apportionment
Description: This short-term effort demonstrated the equivalence of two well-known problems–the optimal allocation of the fixed overall sample size among L strata under stratified random sampling and the optimal allocation of the H = 435 seats among the 50 states for the apportionment of the U.S. House of Representatives following each decennial census. This project continues development with new sample allocation algorithms.

Highlights: Staff completed the draft of a paper on exact optimal sample allocation which presents five different algorithms with various properties. The algorithms all follow from a simple decomposition of sampling error.

Staff: Tommy Wright (x31702), Andrew Perry, Adam Maidman

E. Interviewer-Respondent Interactions: Gaining Cooperation
Description: Survey nonresponse rates have been increasing, leading to concerns about the accuracy of (demographic) sample survey estimates. For example, from 1990 to 2004, initial contact nonresponse rates have approximately doubled for selected household sample surveys including the Current Population Survey (CPS) (from 5.7 percent to 10.1 percent). While mailout/mailback is a relatively inexpensive data collection methodology, decreases in mailback rates to censuses and sample surveys mean increased use of methodologies that bring respondents into direct contact with Census Bureau interviewers (e.g., field representatives) using CATI (computer assisted telephone interviewing) or CAPI (computer assisted

personal interviewing). CAPI can include face-to-face or telephone contact. Unsuccessful interviewer-respondent interactions can lead to increased costs due to the need for additional follow-up, and can also decrease data quality. So, unsuccessful interviewer-respondent interactions should be minimized.

This project will analyze data from 512 field representatives (interviewers) as part of an exploratory study, examining their beliefs regarding what works in gaining respondents' cooperation and investigating associations with field representatives' performance in terms of completed interview rates. We will also study associations between field representatives' beliefs and what they say they do.

Highlights: No significant updates this quarter.

Staff: Tommy Wright (x31702), Tom Petkunas

F. Weighted Estimating Equations with Response Propensities in Terms of Covariates Observed Only for Responders

Description: The project now considers not only survey response propensities but also data on an administrative-records observational database, with the goal of modeling joint indicators of survey response and administrative-list inclusion. Staff aims to develop survey analysis methods incorporating administrative data that can, under suitable model assumptions, provide representative population estimators.

Highlights: During Q3, staff developed an analytical framework and model for joint survey response and administrative-list inclusion, drawing on a literature review conducted this quarter in preparation for a JSM 2016 presentation. Staff prepared a preliminary paper on the current research.

Staff: Eric Slud (x34991), Robert Ashmead, Anindya Roy

G. Analysis and Estimation of Daily Response Propensities and Use of Contact History Instrument (CHI)

Description: Staff will conduct general research methodology to work on existing files to improve modeling accuracy and to provide suggestions for developing and using response propensities. To help the research, we make use of National Crime Victimization Survey data. Staff is also currently using general research methodology to work on simulation study to describe how to reproduce generalized boosted regression modeling algorithm for estimating propensity scores with bootstrap and continuous treatment methods.

Highlights: During Q3 of FY 2016, staff developed methods to fit and evaluate models that can predict daily response propensities. Staff updated existing methodology to describe how to fit daily response propensities along with actual survey indicators and survey outcomes to (1) evaluate model accuracy and determine whether the models need refinement; (2) investigate relationship between response propensity and key survey variables; and (3) determine how the daily response propensities may be used to manage fieldwork.

A formal report of the analysis and estimation documenting work on NCVS daily response propensity modeling and methodology for FY 2016 is currently in progress and nearing completion.

Staff would like to show how this response propensity methodology can provide potential intervention strategy for survey efforts to increase response rates and how more interview cases can result in completions.

Staff: Isaac Dompreh (x36801), Joseph Schafer (ADRM)

Time Series and Seasonal Adjustment

Motivation: Seasonal adjustment is vital to the effective presentation of data collected from monthly and quarterly economic surveys by the Census Bureau and by other statistical agencies around the world. As the developer of the X-13ARIMA-SEATS Seasonal Adjustment Program, which has become a world standard, it is important for the Census Bureau to maintain an ongoing program of research related to seasonal adjustment methods and diagnostics, in order to keep X-13ARIMA-SEATS up-to-date and to improve how seasonal adjustment is done at the Census Bureau.

Research Problems:

- All contemporary seasonal adjustment programs of interest depend heavily on time series models for trading day and calendar effect estimation, for modeling abrupt changes in the trend, for providing required forecasts, and, in some cases, for the seasonal adjustment calculations. Better methods are needed for automatic model selection, for detection of inadequate models, and for assessing the uncertainty in modeling results due to model selection, outlier identification and non-normality. Also, new models are needed for complex holiday and calendar effects.
- Better diagnostics and measures of estimation and adjustment quality are needed, especially for model-based seasonal adjustment.
- For the seasonal, trading day and holiday adjustment of short time series, meaning series of length five years or less, more research into the properties of methods usually used for longer series, and perhaps into new methods, are needed.

Potential Applications:

- To the effective presentation of data collected

from monthly and quarterly economic surveys by the Census Bureau and by other statistical agencies around the world.

A. Seasonal Adjustment
Description: This research is concerned with improvements to the general understanding of seasonal adjustment and signal extraction, with the goal of maintaining, expanding, and nurturing expertise in this topic at the Census Bureau.

Highlights: Staff (a) completed methodology, testing, and discussion of a maximum entropy extreme value adjustment of New Zealand agricultural time series; (b) continued empirical work on seasonal heteroscedasticity models to show improved forecasting and seasonal adjustment of construction series; and (c) extended work on signal extraction decompositions allowing for correlation between components.

Staff: Tucker McElroy (x33227), James Livsey, Brian Monsell, Osbert Pang, Anindya Roy

B. Time Series Analysis
Description: This research is concerned with broad contributions to the theory and understanding of discrete and continuous time series, for univariate or multivariate time series. The goal is to maintain and expand expertise in this topic at the Census Bureau.

Highlights: During Q3 of FY 2016, staff (a) continued work on stable parametrizations of VARMA models fitted under parameter constraints, utilizing a LASSO objective function; (b) completed study of the multivariate bullwhip effect for retail supply chains; (c) implemented and tested likelihood ratio tests for Granger non-causality, as a way to exclude extraneous data from multivariate forecasting problems; (d) continued research and simulations for two tests of co-integration, one based upon fitted structural models and another based on nonparametric spectral estimates; (e) continued research, development, and testing for method of assessing the entropy of model residuals; (f) conducted methodological work on high-dimensional time series signal extraction, for possible use on space-time data; (g) continued research on the Frobenius norm, as a tool for fitting and evaluating multivariate time series models; and (h) continued implementation and development of vector band pass and low pass filters.

Staff: Tucker McElroy (x33227), David Findley (Private Collaborator), Brian Monsell, James Livsey, Osbert Pang, Anindya Roy

C. Time Series Model Development
Description: This work develops a flexible integer-valued autoregressive (AR) model for count data that contain data over- or under-dispersion (i.e. count data where the variance is larger or smaller than the mean, respectively). Such a model will contain Poisson and Bernoulli AR models as special cases.

Highlights: No significant updates this quarter.

Staff: Kimberly Sellers (x39808)

Experimentation and Statistical Modeling

Motivation: Experiments at the Census Bureau are used to answer many research questions, especially those related to testing, evaluating, and advancing survey sampling methods. A properly designed experiment provides a valid, cost-effective framework that ensures the right type of data is collected as well as sufficient sample sizes and power are attained to address the questions of interest. The use of valid statistical models is vital to both the analysis of results from designed experiments and in characterizing relationships between variables in the vast data sources available to the Census Bureau. Statistical modeling is an essential component for wisely integrating data from previous sources (e.g., censuses, sample surveys, and administrative records) in order to maximize the information that they can provide.

Research Problems:
• Investigate bootstrap methodology for sample surveys; implement the bootstrap under complex sample survey designs; investigate variance estimation for linear and non-linear statistics and confidence interval computation; incorporate survey weights in the bootstrap; investigate imputation and the bootstrap under various non-response mechanisms.
• Investigate methodology for experimental designs embedded in sample surveys; investigation of large-scale field experiments embedded in ongoing surveys; design based and model based analysis and variance estimation incorporating the sampling design and the experimental design; factorial designs embedded in sample surveys and the estimation of interactions; testing non-response using embedded experiments. Use simulation studies.
• Assess feasibility of established design methods (e.g., factorial designs) in Census Bureau experimental tests.
• Identify and develop statistical models (e.g., loglinear models, mixture models, and mixed-effects models) to characterize relationships between variables measured in censuses, sample surveys, and administrative records.
• Assess the applicability of post hoc methods (e.g., multiple comparisons and tolerance intervals) with future designed experiments and when reviewing previous data analyses.

Potential Applications:

• Modeling approaches with administrative records can help enhance the information obtained from various sample surveys.
• Experimental design can help guide and validate testing procedures proposed for the 2020 Census.
• Expanding the collection of experimental design procedures currently utilized with the American Community Survey.

A. Design and Analysis of Embedded Experiments
Description: This ongoing project will explore rigorous analysis of embedded experiments: from simple idealized designs to complex designs used in practice at the Census Bureau.

Highlights: Staff investigated variance estimation in the setting where two treatments are compared in a sample taken from a finite population. A naïve bootstrap estimator was seen to be nearly equivalent to an estimator proposed by Van den Brakel & Renssen (VDBR, 1996). Staff evaluated more sophisticated bootstrap and permutation methods using several sampling designs. Staff found the VDBR estimator to have the smallest bias, but other methods may have smaller variability under complex sampling designs. This is currently under investigation.

Staff: Thomas Mathew (x35337), Andrew Raim

B. Synthetic Survey and Processing Experiments
Description: To improve operational efficiencies and reduce costs of survey processing, this project will simulate a survey, in which an artificial team of interviewers seek out an artificial set of respondents, to test alternative methods of allocating resources in the field and to test alternatives for the post-processing of the gathered survey data. When calibrated with survey paradata, the model may also serve as a test bed for new methods of missing data imputation.

Highlights: This project is currently on hold.

Staff: TBA

C. Multivariate Nonparametric Tolerance Regions
Description: A tolerance region for a multivariate population is a region computed using a random sample that will contain a specified proportion or more of the population, with a given confidence level. Typically, tolerance regions that have been computed for multivariate populations are elliptical in shape. A difficulty with an elliptical region is that it cannot provide information on the individual components of the measurement vector. However, such information can be obtained if we compute tolerance regions that are rectangular in shape. This project applies bootstrap ideas to compute multivariate tolerance regions in a nonparametric framework. Such an approach can be applied to multivariate economic data and aid in the editing process by identifying multivariate observations that are outlying in one or more attributes and subsequently should undergo further review.

Highlights: Significant progress has been made on developing the necessary theoretical framework. The approach consists of trimming the multivariate data set by employing statistical data depth, and utilizing the extremes of the trimmed dataset as the faces of the hyper-rectangular region. A strategy for determining the number of points to be trimmed is developed, and an algorithm is provided for implementing the methodology. An extensive coverage study shows the favorable performance of the algorithm for moderate to large sample sizes. For smaller sample sizes, a bootstrap calibration routine is recommended for improved performance. A manuscript based on the work is under preparation.

Staff: Thomas Mathew (x35337)

D. Master Address File (MAF) Research—Developing a Generalized Regression Model for Count Data
Description: This project develops a zero-inflated version of a generalized regression model for count data based on the Conway-Maxwell-Poisson distribution to allow for data-dispersion and excess zeroes in the dataset. The objective of this project is to develop and consider an alternative regression model for use to describe associations with changes in the number of housing units (adds or deletes) on a block, and predict where housing growth or decline may occur in the MAF.

Highlights: This project is now complete.

Staff: Kimberly Sellers (x39808), Andrew Raim

E. Modeling the Causal Effects of Field Representative Actions and Strategies
Description: Field Representatives (FRs) apply different strategies for managing their monthly workloads. For example, some FRs may place high priority on contacting households that are perceived as likely to respond, putting aside the more difficult cases until later in the month. With large volumes of information flowing from paradata systems, we are better able to model FR data collection behavior. However, to understand the causal effects of these behaviors on outcomes of interest (response rates, measures of data quality and measures of cost), we need to adjust for confounding characteristics of FRs and their caseloads. In this project, we are developing techniques for causal inference from observational (non-experimental) data on FR characteristics, behaviors and performance measures.

Highlights: No significant updates this quarter.

Staff: Doug Galagate (x34985), Robert Ashmead

F. Development of a Bivariate Distribution for Count Data where Data Dispersion is Present

Description: This project develops a bivariate form of the Conway-Maxwell-Poisson distribution to serve as a tool to describe variation and association for two count variables that express over- or under-dispersion (relationships where the variance of the data is larger or smaller than the mean, respectively).

Highlights: A manuscript associated with this work was published in the *Journal of Multivariate Analysis.*

Staff: Kimberly Sellers (x39808), Darcy Steeg Morris

G. Developing a Flexible Stochastic Process for Significantly Dispersed Count Data

Description: The Bernoulli and Poisson are two popular count processes; however, both rely on strict assumptions that motivate their use. CSRM staff (with other collaborators) instead propose a generalized count process (hereafter named the Conway-Maxwell-Poisson process) that not only includes the Bernoulli and Poisson processes as special cases, but also serves as a flexible mechanism to describe count processes that approximate data with over- or under-dispersion. Staff introduce the process and its associated generalized waiting time distribution with several real-data applications to illustrate its flexibility for a variety of data structures. This new generalized process will enable analysts to better model count processes where data dispersion exists in a more accommodating and flexible manner.

Highlights: No significant updates this quarter.

Staff: Kimberly Sellers (x39808), Darcy Steeg Morris

H. Analysis of Under-dispersed Count Data

Description: This research concerns contributions to the theory and understanding of under-dispersed count data, and models that accommodate such data. The goal is to expand understanding and expertise in this area at the Census Bureau.

Highlights: Staff have begun to study data under-dispersion. Staff completed a literature review of the topic, addressing causes of data under-dispersion and learning of various statistical models that either accommodate only for under-dispersion, or over- or under-dispersion. Initial research and results will be presented at an invited session at the XXVIIIth International Biometric Conference in Victoria, BC, Canada.

Staff: Kimberly Sellers (x39808), Darcy Morris

Simulation and Statistical Modeling

Motivation: Simulation studies that are carefully designed under realistic survey conditions can be used to evaluate the quality of new statistical methodology for Census Bureau data. Furthermore, new computationally intensive statistical methodology is often beneficial because it can require less strict assumptions, offer more flexibility in sampling or modeling, accommodate complex features in the data, enable valid inference where other methods might fail, etc. Statistical modeling is at the core of the design of realistic simulation studies and the development of intensive computational statistical methods. Modeling also enables one to efficiently use all available information when producing estimates. Such studies can benefit from software such as *Tea* for data processing. Statistical disclosure avoidance methods are also developed and properties studied.

Research Problems:

- Systematically develop an environment for simulating complex surveys that can by used as a test-bed for new data analysis methods.
- Develop flexible model-based estimation methods for survey data.
- Develop new methods for statistical disclosure control that simultaneously protect confidential data from disclosure while enabling valid inferences to be drawn on relevant population parameters.
- Investigate the bootstrap for analyzing data from complex sample surveys.
- Continue to formalize the codebase and user interfacing for *Tea*, especially within the context of the current enterprise environment.
- Develop models for the analysis of measurement errors in Demographic sample surveys (e.g., Current Population Survey or the Survey of Income and Program Participation).
- Identify and develop statistical models (e.g., loglinear models, mixture models, and mixed-effects models) to characterize relationships between variables measured in censuses, sample surveys, and administrative records.
- Investigate noise multiplication for statistical disclosure control.

Potential Applications:

- Simulating data collection operations using Monte Carlo techniques can help the Census Bureau make more efficient changes.
- Use noise multiplication or synthetic data as an alternative to top coding for statistical disclosure control in publicly released data. Both noise multiplication and synthetic data have the potential to preserve more

information in the released data over top coding.
• Rigorous statistical disclosure control methods allow for the release of new microdata products.
• *Tea* provides modeling and editing flexibility, especially with a focus on incorporating administrative data.
• Using an environment for simulating complex surveys, statistical properties of new methods for missing data imputation, model-based estimation, small area estimation, etc. can be evaluated.
• Model-based estimation procedures enable efficient use of auxiliary information (for example, Economic Census information in business surveys), and can be applied in situations where variables are highly skewed and sample sizes are not sufficiently large to justify normal approximations. These methods may also be applicable to analyze data arising from a mechanism other than random sampling.
• Variance estimates and confidence intervals in complex surveys can be obtained via the bootstrap.
• Modeling approaches with administrative records can help enhance the information obtained from various sample surveys.

A. Development and Evaluation of Methodology for Statistical Disclosure Control
Description: When survey organizations release data to the public, a major concern is the protection of individual records from disclosure while maintaining quality and utility of the released data. Procedures that deliberately alter data prior to their release fall under the general heading of statistical disclosure control. This project develops new methodology for statistical disclosure control, and evaluates properties of new and existing methods. We develop and study methods that yield valid statistical analyses, while simultaneously protecting individual records from disclosure.

Highlights: Staff continued studying inference based on synthetic data when the data generating model, imputation model, and data analysis model are not all the same. Staff specifically considered synthetic data created using posterior predictive sampling under a linear regression model, and theoretically evaluated the effects of underfitting or overfitting the imputation model and/or data analysis model. This evaluation was done for both the case of singly imputed synthetic data as well as multiply imputed synthetic data.

Staff: Martin Klein (x37856), Bimal Sinha (CDAR), Thomas Mathew, Brett Moran

Summer at Census

Description: For each summer since 2009, recognized scholars in the following and related fields applicable to censuses and large-scale sample surveys are invited for short-term visits (one to five days) primarily between May and September: statistics, survey methodology, demography, economics, geography, social and behavioral sciences, and computer science. Scholars present a seminar based on their research and engage in collaborative research with Census Bureau researchers and staff.

Scholars are identified through an annual Census Bureau-wide solicitation by the Center for Statistical Research and Methodology.

Highlights: Staff facilitated all the details and background with staff from around the Census Bureau to host *2016 SUMMER AT CENSUS* with nearly forty scholors.

Staff: Tommy Wright (x31702), Michael Leibert

Research Support and Assistance

This staff provides substantive support in the conduct of research, research assistance, technical assistance, and secretarial support for the various research efforts.

Staff: Erica Magruder, Kelly Taylor

3. PUBLICATIONS

3.1 JOURNAL ARTICLES, PUBLICATIONS

Lu, X. and West, D. (2016). "A New Proof that 4-connected Planar Graphs are Hamiltonian-connected," *Discussiones Mathematicae Graph Theory, 36:* 555-564.

Blakely, C. and McElroy, T. (2016). "Signal Extraction Goodness-of-fit Diagnostic Tests Under Model Parameter Uncertainty," *Econometrics Reviews,* 1-16.

McElroy, T. (2016). "Multivariate Seasonal Adjustment, Economic Identities, and Seasonal Taxonomy," *Journal of Business and Economics Statistics.* Published Online.

McElroy, T. (2016). "On the Measurement and Treatment of Extremes in Time Series," *Extremes,* 1-24.

McElroy, T. and McCracken, M. (2016). "Multi-Step Ahead Forecasting of Vector Time Series," *Econometrics Reviews,* 1-26.

Morris, D.S., Keller, A., and Clark, B. (2016). "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census." *Statistical Journal of the International Association for Official Statistics, 32(2)*: 177-178.

Mulry, Mary H., Nichols, Elizabeth M. & Childs Hunter, J. (2016). "A Case Study of Error in Survey Reports of Move Month Using the U.S. Postal Service Change of Address Records," *Survey Methods: Insights from the Field.* Retrieved from http://surveyinsights.org/?p=7794.

Sellers, K.F., Morris, D.S., and Balakrishnan, N. (2016). "Bivariate Conway-Maxwell-Poisson Distribution: Formulation, Properties, and Inference," *Journal of Multivariate Analysis, 150*: 152-168.

Wildi, M. and McElroy, T. (2016). "Optimal Real-Time Filters for Linear Prediction Problems," *Journal of Time Series Econometrics, 8:155-192*.

3.2 BOOKS/BOOK CHAPTERS

3.3 PROCEEDINGS PAPERS

3.4 CENTER FOR STATISTICAL RESEARCH & METHODOLOGY RESEARCH REPORTS
<http://www.census.gov/srd/www/byyear.html>

RR (Statistics #2016-03): Tommy Wright. "Two Optimal Exact Sample Allocation Algorithms: Sampling Variance Decomposition is Key," May 10, 2016.

3.5 OTHER REPORTS

4. TALKS AND PRESENTATIONS

Work Outside the Book: Humanities Career Tracks Outside of Academia, University of Maryland, College Park, Maryland, April 6, 2016.

- Lauren Emanuel, "English Majors in the Federal Government."

SAMSI Workshop on Games and Decisions in Reliability and Risk, Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, North Carolina, May 18, 2016.

- Kimberly Sellers, "A Generalized Statistical Control Chart for Over- or Under-dispersed Data."

10th Annual Probability & Statistics Day at UMBC, Baltimore County, Maryland, May 20-21, 2016.

- Martin Klein, "Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multiple Linear Regression Model."
- Tommy Wright, "Measurement for Official Statistics."

Department of Biostatistics Seminar, University of Iowa, Iowa City, Iowa, June 17, 2016.

- Tommy Wright, "The Equivalence of Neyman Optimum Allocation for Sampling and Equal Proportions for Apportioning the U.S. House of Representatives."

Iowa Summer Institute in Biostatistics, University of Iowa, Iowa City, Iowa, June 17, 2016.

- Tommy Wright, "Measurement and Official Statistics."

5. CENTER FOR STATISTICAL RESEARCH AND METHODOLOGY SEMINAR SERIES

Yves Thibaudeau and William Winkler, U.S. Census Bureau, "Edit & Imputation Theory & Computational Algorithms," April 6, 13, & 20, 2016.

Douglas Galagate, U.S. Census Bureau/University of Maryland, College Park, "Causal Inference with a Continuous Treatment and Outcome: Alternative Estimators for Parametric Dose-Response Functions," May 10, 2016.

Yinglei Lai, The George Washington University, *SUMMER AT CENSUS*, "Exploration of Concordant Changes among Multiple Data Sets," May 17, 2016.

Subhashis Ghoshal, North Carolina State University, *SUMMER AT CENSUS*, "Bayesian Nonparametric Methods for Data-Science," May 18, 2016.

Randall Akee, University of California, Los Angeles, *SUMMER AT CENSUS*, "Land Titles and Dispossession: Allotment on American Indian Reservations," May 24, 2016.

Philip L.H. Yu, The University of Hong Kong, *SUMMER AT CENSUS*, "Rank Aggregation Using Distance-Based Models," May 25, 2016.

Domingo Morales, University of Miguel Hernández of Elche, Spain, *SUMMER AT CENSUS*, "Multivariate Fay-Herriot Models for Small Area Estimation," May 31, 2016.

Domingo Morales, University of Miguel Hernández of Elche, Spain, *SUMMER AT CENSUS*, "Small Area Estimation of Non-Linear Parameters Under a Two-Fold Nested Error Regression Model," June 1, 2016.

Jae-Kwang Kim (ASA/NSF/Census Research Fellow), Iowa State University, "Some Recent Topics on Informative Sampling," June 2, 2016.

James D. Wilson, University of San Francisco, "A Significance-based Community Extraction Method for Multilayer Networks," June 2, 2016.

Zachary Seeskin, (U.S. Census Bureau Dissertation Fellow), Northwestern University, "Evaluating the Use of Commercial Data to Improve Survey Estimates of Property Taxes," June 7, 2016.

John Iceland, The Pennsylvania State University, *SUMMER AT CENSUS*, "Did We Win the War on Poverty? No, but…," June 7, 2016.

Wendy Manning, Bowling Green State University, *SUMMER AT CENSUS*, "Measuring Cohabitation in National Surveys," June 8, 2016.

Sheela Kennedy, University of Michigan, *SUMMER AT CENSUS*, "The Changing Transition to Adulthood in the U.S.: Trends in Demographic Role Transitions and Age Norms since 2000," June 9, 2016.

Sharon Sassler, Cornell University, *SUMMER AT CENSUS*, "A Cross-National Comparison of the Consequences of Partnered Childbearing for Mother's Mid-Life Health," June 9, 2016.

Vitaly Shmatikov, Cornell Tech, *SUMMER AT CENSUS*, "Machine Learning and Privacy: Friends or Foes?," June 14, 2016.

Henry Schneider, Cornell University, *SUMMER AT CENSUS*, "Promoting Best Practices in a Multitask Workplace: Experimental Evidence on Checklists," June 14, 2016.

Carolyn Liebler, University of Minnesota, *SUMMER AT CENSUS*, "The Occupational Structure of the American Indian and Alaska Native Workforce," June 20, 2016.

Christoph Sax, Christoph Sax Data Analytics LLC, *SUMMER AT CENSUS*, "R-Development: User Interfaces and Package Creation," June 21, 2016.

Zhuoqiong He, University of Missouri-Columbia, *SUMMER AT CENSUS*, "Assessing and Adjusting Nonresponse Bias in Small Area Estimation via Bayesian Hierarchical Spatial Models," June 21, 2016.

Elizabeth Fussell, Brown University, *SUMMER AT CENSUS*, "Disasters and Residential Change in the U.S., 1997-2013: Migrants' Reasons for Moving, Socio-Demographic Selection, and Residential Outcomes," June 27, 2016.

Ashley Amaya (U.S. Census Bureau Dissertation Fellow, University of Maryland), RTI International, "Enhancing the Understanding of the Relationship between Social Integration and Nonresponse," June 28, 2016.

Narayan Sastry, University of Michigan, *SUMMER AT CENSUS*, "The Effects of Hurricane Katrina on the New Orleans Population: Results from the American Community Survey," June 28, 2016.

Scott Holan, University of Missouri, *SUMMER AT CENSUS*, "Multivariate Spatio-Temporal Models for High-Dimensional Areal Data with Application to Longitudinal Employer-Household Dynamics," June 29, 2016.

6. PERSONNEL ITEMS

6.1 HONORS/AWARDS/SPECIAL RECOGNITION

6.2 SIGNIFICANT SERVICE TO PROFESSION

Emanuel Ben-David

- Refereed papers for *Mathematical Reviews, Statistica Sinica,* and *Journal of Statistical Planning*

Carolina Franco

- Refereed papers for *The American Statistician* and the *Journal of Official Statistics*

Martin Klein

- Refereed a paper for *Journal of the Royal Statistical Society-Series A*
- Member, Ph.D. Dissertation in Statistics Committee, University of Maryland, Baltimore County
- Chaired a session at the 10th Annual Probability & Statistics Day at University of Maryland, Baltimore County

Thomas Mathew

- Associate Editor, *Journal of the American Statistical Association*
- Associate Editor, *Statistical Methodology*
- Associate Editor, *Sankhya, Series B*
- Editorial Board member, *Journal of Occupational and Environmental Hygiene*
- Member, American Statistical Association's Committee on W.J. Youden Award in Inter-laboratory Testing
- Reviewed articles for *Journal of the American Statistical Association, Journal of Official Statistics, Sankhya, Statistics and Probability Letters* and *Communications in Statistics*

Kimberly Sellers

- Member, American Statistical Association Committee on Women in Statistics
- Associate Editor, *The American Statistician*
- Advisory Board Member and Director, BDN STEMers for International Black Doctoral Network Association, Incorporated
- Refereed papers for *Applied Stochastic Models in Business and Industry, Biometrics, Communications in Statistics – Theory and Methods, Computers & Industrial Engineering, Lifetime Data Analysis, Quality and Reliability Engineering International,* and *Statistics*
- Member, Scientific Program Committee, International Conference on Statistical Distributions and Applications (ICOSDA) 2016

William Winkler

- Refereed papers for *JASA* and *Statistical Data Protection 2016*
- Reviewed a grant proposal on record linkage for the National Science Foundation
- Associate Editor, *Journal of Privacy and Confidentiality*
- Associate Editor, *Transactions on Data Privacy*
- Member, Program Committee for *Statistical Data Protection 2016*
- Member, Program Committee for *IEEE 2016 ICDM Data Integration and Applications*
- Member, Statistics Ph.D. Committee at the University of Maryland

Tommy Wright

- Associate Editor, *The American Statistician*
- Chair, Waksberg Award Committee, *Survey Methodology*
- Member, Board of Trustees, National Institute of Statistical Sciences

6.3 PERSONNEL NOTES

Rolando Rodriguez accepted a position in the Center for Disclosure Avoidance.

Bret Hanlon joined our Simulation, Modeling, and Data Visualization Research Group.

Adam Maidman (Ph.D student in statistics at University of Minnesota) joined our center as a summer intern.

Jae-Kwang Kim (Statistics Professor at Iowa State University) joined the Census Bureau as an ASA/NSF/Census Research Fellow.